Transformers for jet identification in particle physics

Congqiao Li (李聪乔), ***Peking University***

量子计算和人工智能与高能物理交叉研讨会 · 中国科学技术大学

11 January, 2025

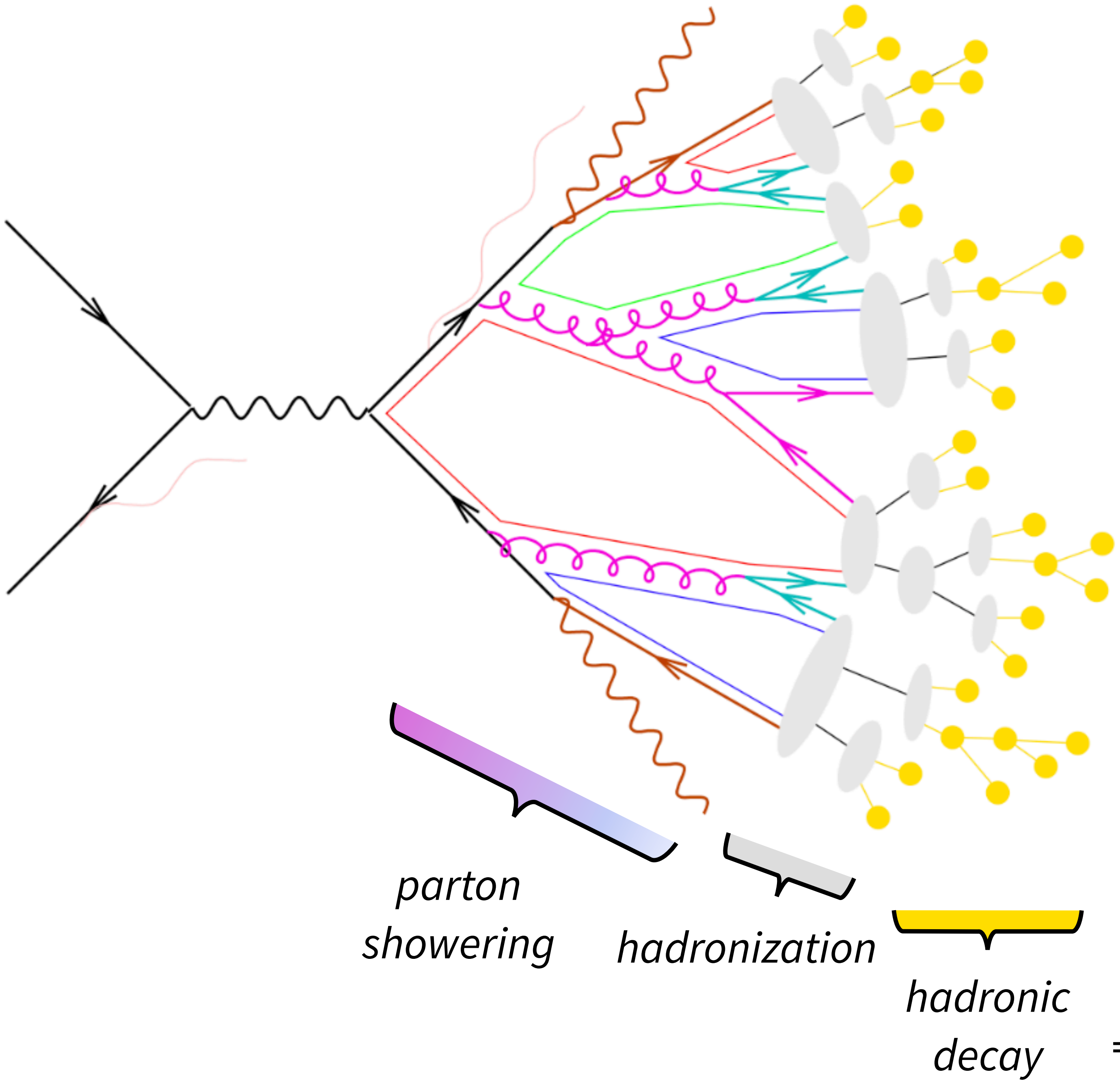
Outline

This talk will cover

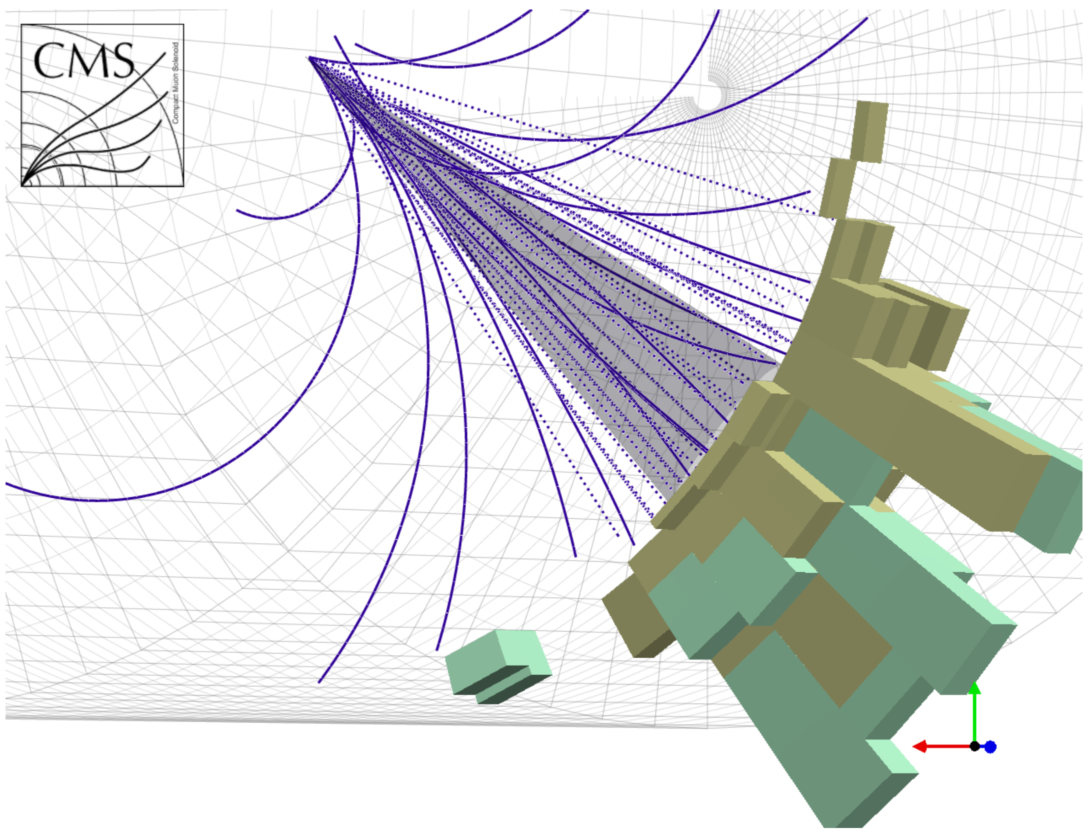
- ➔ Evolution of DNNs for jet identification
  - ❖ a deep overview of gained experiences from the prior developments
- ➔ Transformer models for jets
  - ❖ how to adapt Transformer networks to jet physics?
  - ❖ advances & application examples
  - ❖ future insights

Jets in particle physics

Jets are collinear sprays of particles initiated by quark/gluons



⇒ *stable hadrons*

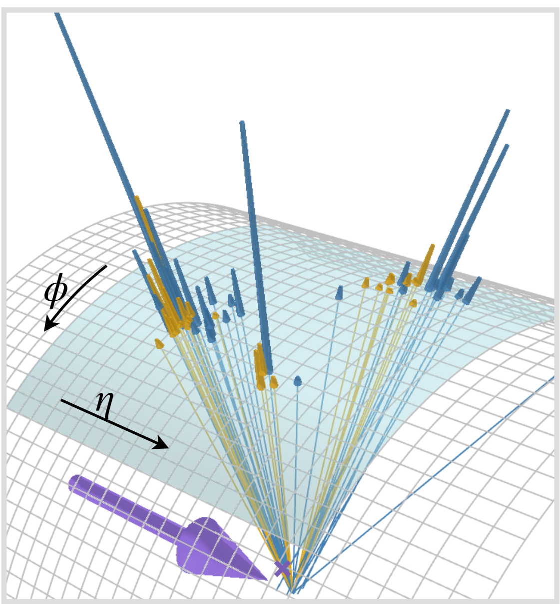


raw data from tracker & calorimeter
→ reconstruct to particle records (particle-flow candidates in CMS) to cluster jets

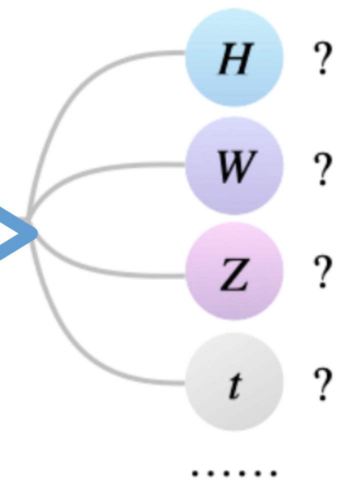
Jet identification (jet tagging): identify the origin of the jet

Question: How to design a most performant jet NN?

➔ This is a highly physics-ML interdisciplinary subject



original of the jet?

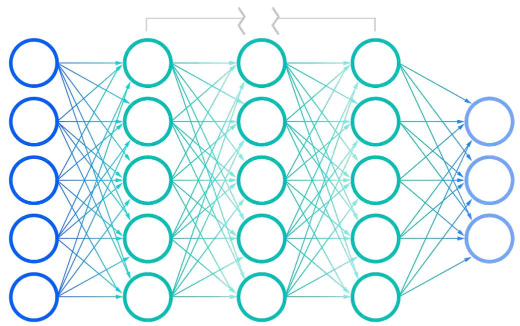


other jet properties?

research objective:
design a most performant jet NN

Evolution of jet NNs

feed-forward NN *(high-level inputs)* ····▶···· *1D/2D CNN, RNN* *(low-level inputs)* ····▶··· *graph NN, Transformers* *(low-level inputs)* ····▶··· *??*

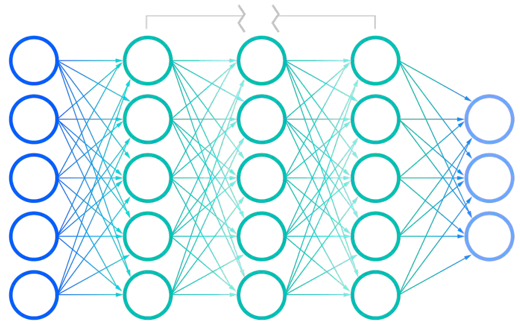


Shallow networks

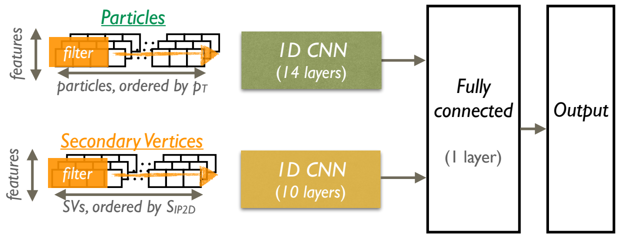
✦ Using high-level features directly as input to a shallow network

Evolution of jet NNs

feed-forward NN *(high-level inputs)* ····▶··· ***1D/2D CNN, RNN*** *(low-level inputs)* ····▶··· *graph NN, Transformers* *(low-level inputs)* ····▶··· *??*



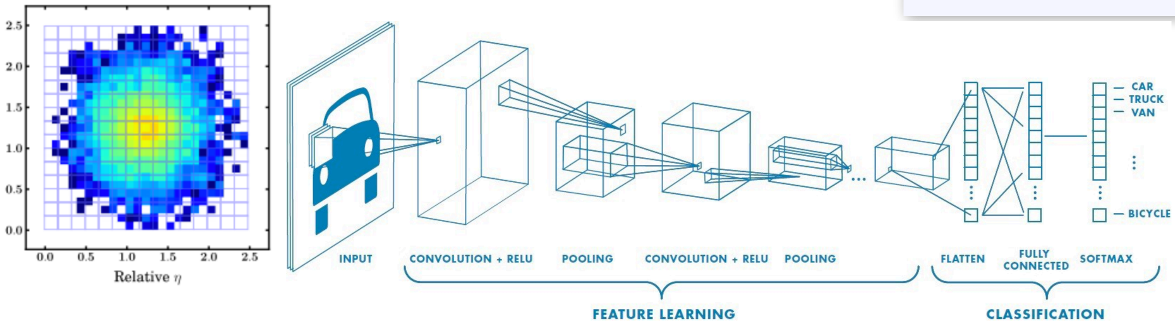
Shallow networks



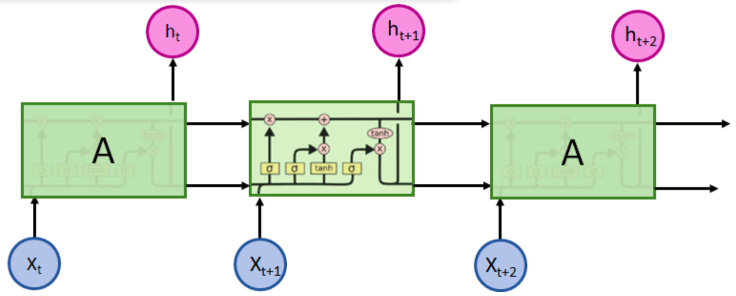
Deep NN with low-level inputs

- ✦ Using high-level features directly as input to a shallow network

- ✦ Using particle-level features
- ✦ Input data structure determines the type of networks
  - jet as a *image (fixed-grid data structure)*
  - jet as a *sequence* → 1D CNN or RNN



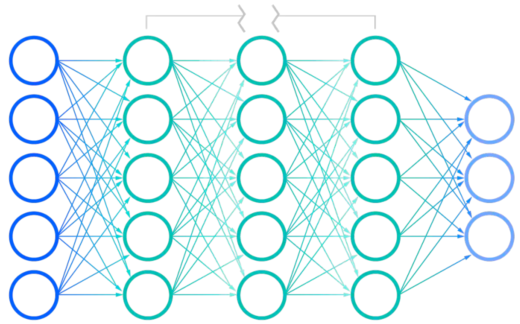
Typical CNN



Typical RNN

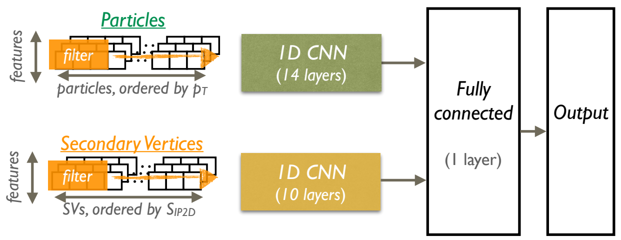
Evolution of jet NNs

feed-forward NN (high-level inputs) ····▶··· **1D/2D CNN, RNN** *(low-level inputs)* ····▶··· *graph NN, Transformers (low-level inputs)* ····▶··· ??



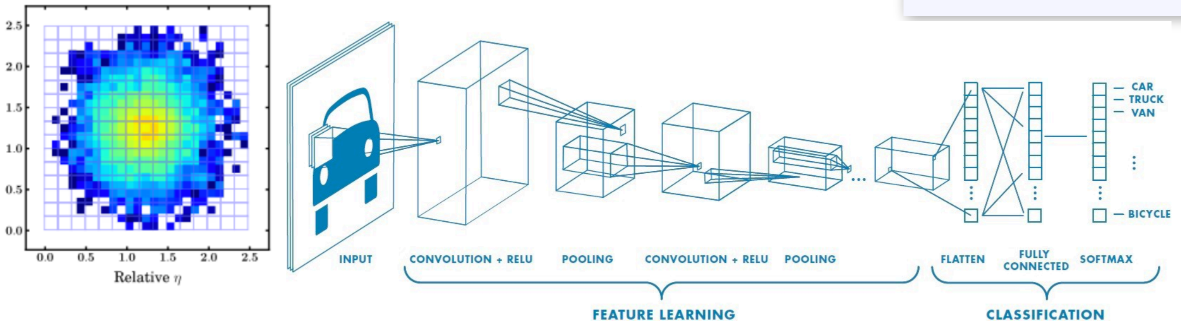
Shallow networks

✦ Using high-level features directly as input to a shallow network



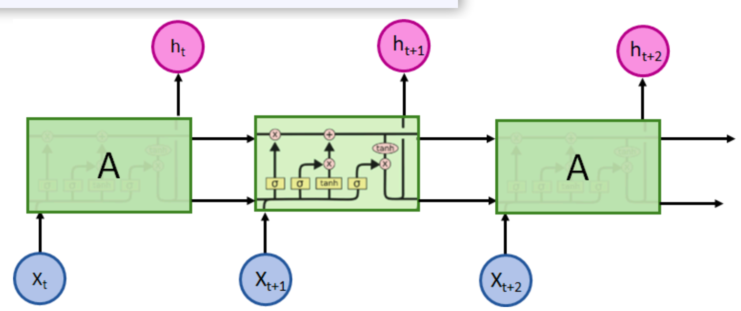
Deep NN with low-level inputs

✦ Using particle-level features
✦ Input data structure determines the type of networks
- jet as a *image (fixed-grid data structure)*
- jet as a *sequence* → 1D CNN or RNN



Typical CNN

deficiency:
has information loss,
brings data sparsity

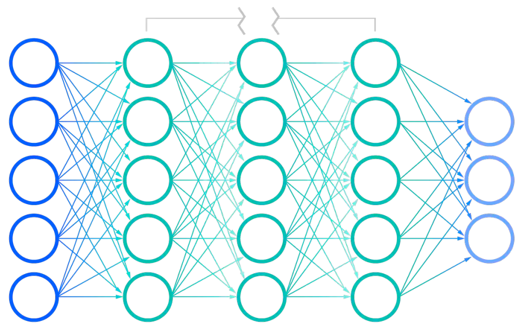


Typical RNN

deficiency:
introduce artificial order
hard to capture long-term dependencies

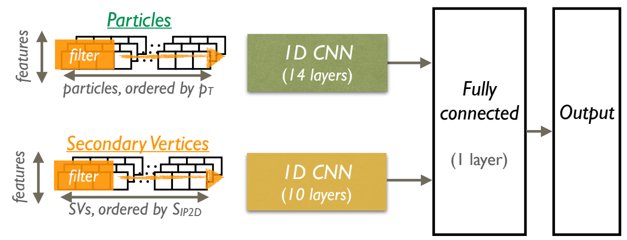
Evolution of jet NNs

feed-forward NN (high-level inputs) ····▶···· *1D/2D CNN, RNN (low-level inputs)* ····▶··· ***graph NN, Transformers*** ····▶··· *??*
(low-level inputs)



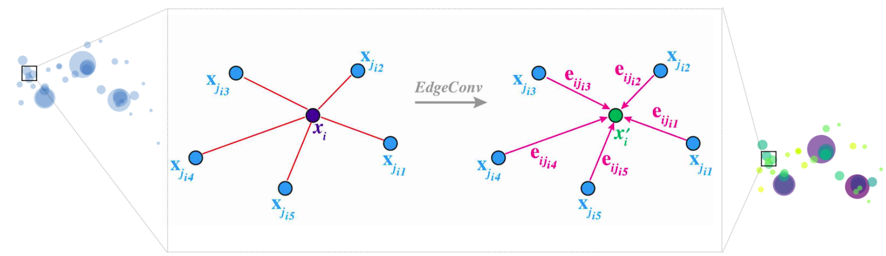
Shallow networks

✦ Using high-level features directly as input to a shallow network



Deep NN with low-level inputs

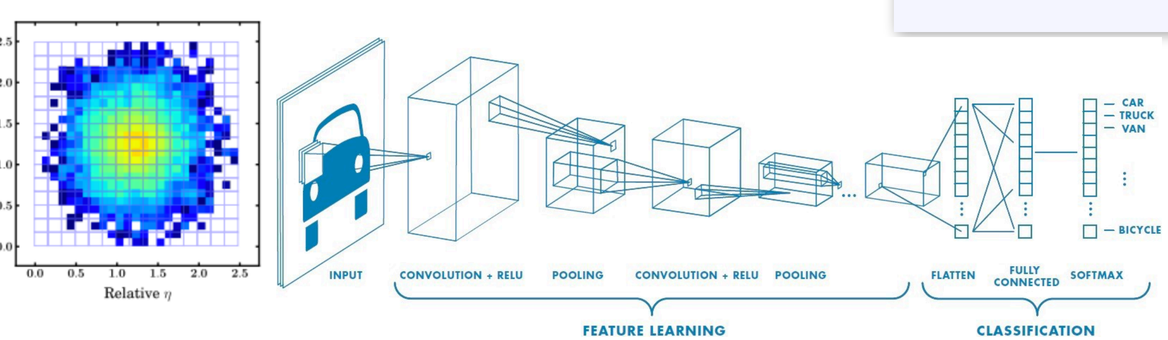
✦ Using particle-level features
✦ Input data structure determines the type of networks
- jet as a *image (fixed-grid data structure)*
- jet as a *sequence* → 1D CNN or RNN



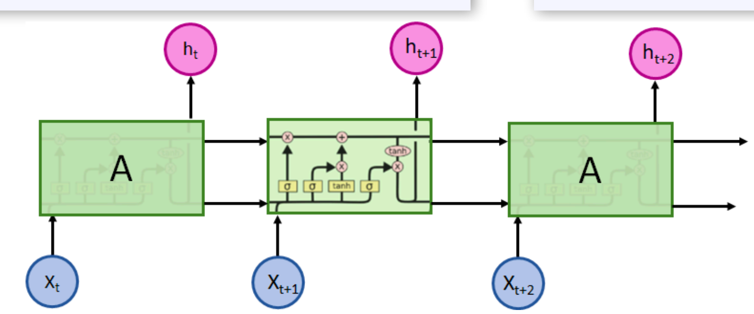
Graph structure

✦ Graph neural networks
- treat a jet as a **permutational-invariant** set of particles (or, point cloud)
- build "edges" between particles

✦ Transformer networks
- modern architectural designs - act like a "fully connected graph"



Typical CNN



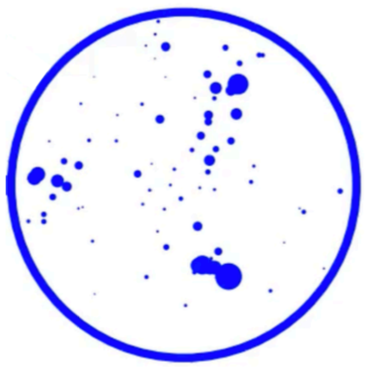
Typical RNN

Typical graph

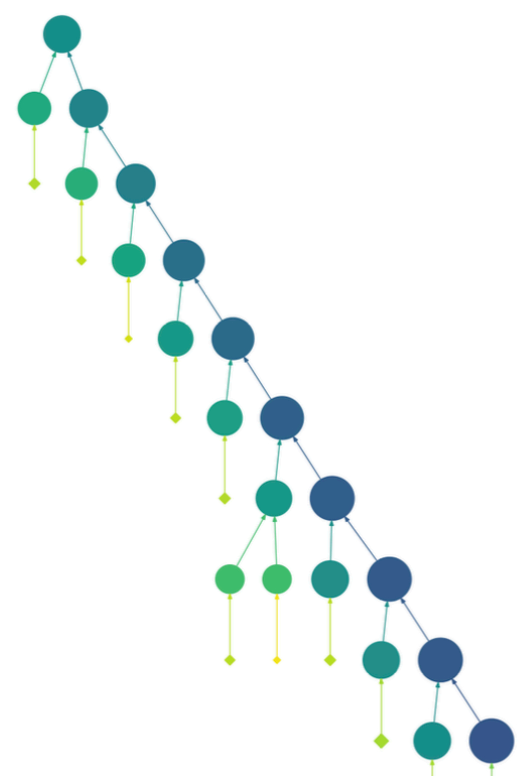
Set/graph representations of jets

- ➔ View input particles as a set/graph
  - ❖ guarantee the *permutational invariance* of input particles
  - ❖ a special stage in jet network developments
- ➔ The **edges** of graph: enable communication between pairs of particles

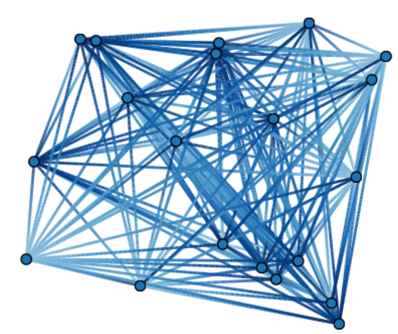
Set: no edges



Hierarchical trees:
- *decay chain*
- *jet clustering history*

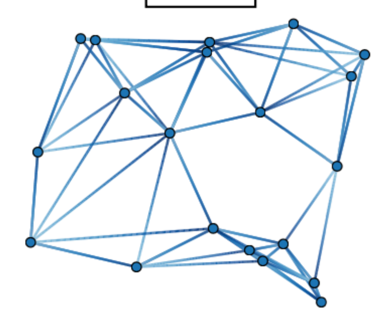


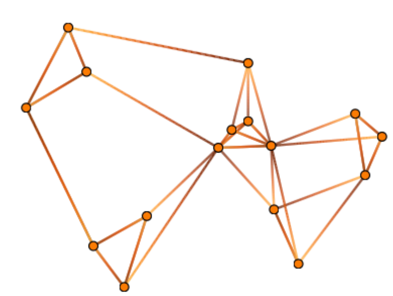
Fully connected graph
- *i.e., connect each node to all other nodes*



Locally connected graph
- *i.e., connect each node only to neighbor nodes*
  - *k-nearest neighbors*
  - *fixed radius*

static





(dynamically) learned

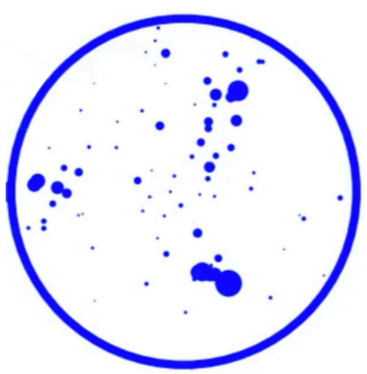
[image from link]

Set/graph representations of jets

- ➔ View input particles as a set/graph
  - ❖ guarantee the *permutational invariance* of input particles
  - ❖ a special stage in jet network de
- ➔ The **edges** of graph: enable com particles

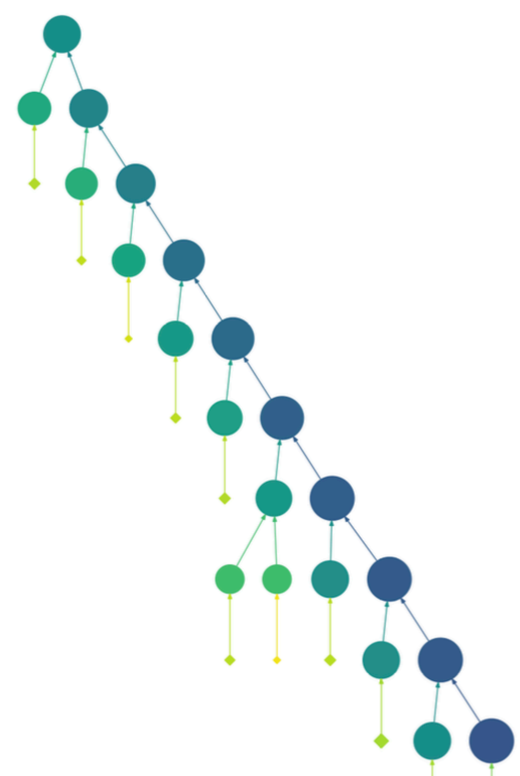
LorentzNet: *S. Gong et al. JHEP 07 (2022) 030*
ParT: *H. Qu et al. arXiv:2202.03772, ICML 2022*
CPT : *S. Qiu et al. PRD 107 (2023) 11, 114029*
HMPNet : *F. Ma et al. PRD 108 (2023) 7, 072007*

Set: no edges



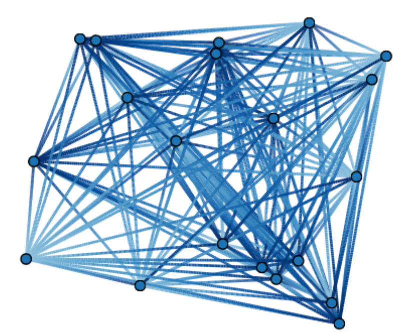
PFN/EFN: *P. Komiske et al. JHEP 01 (2019) 121*

Hierarchical trees:
- *decay chain*
- *jet clustering history*



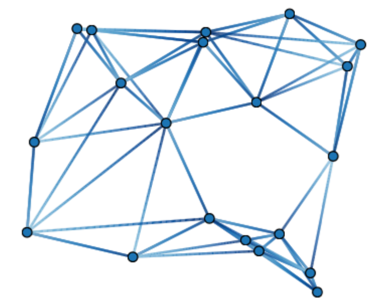
LundNet: *F. Dreyer et al. JHEP 03 (2021) 052*

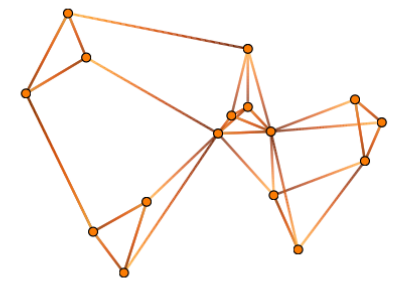
Fully connected graph
- *i.e., connect each node to all other nodes*



Locally connected graph
- *i.e., connect each node only to neighbor nodes*
  - *k-nearest neighbors*
  - *fixed radius*

static





(dynamically) learned

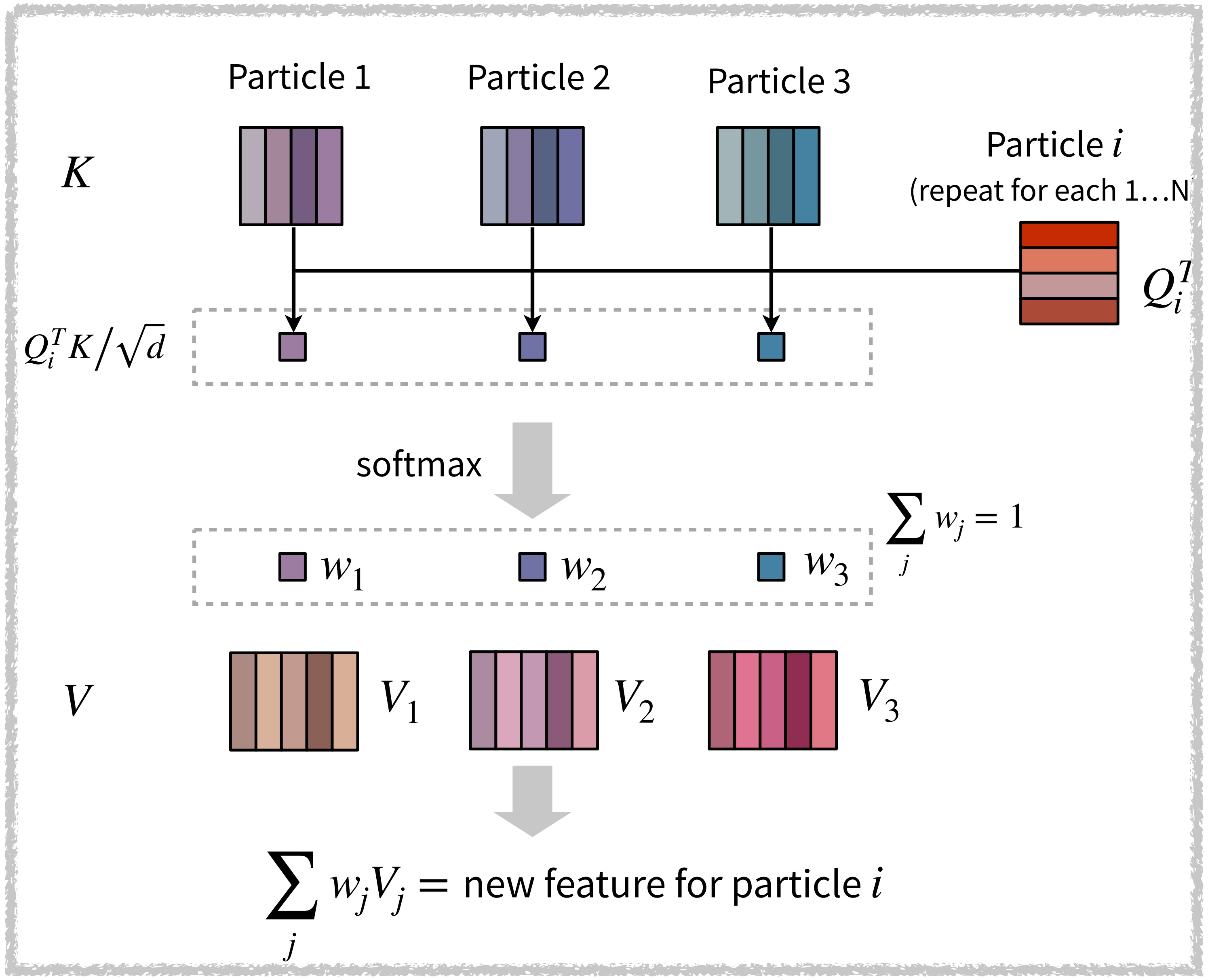
[image from link]

ParticleNet: *H.Qu et al. PRD 101, 056019 (2020)*
ABCNet: *V. Mikuni et al. EPJC 2020; 135(6): 463*

Transformer × jet network?

- Transformer (Google, 2017): unifies the architecture designs across the tasks
  - initiated in NLP, then extended to computer vision (started by ViTs)
- Benefits:
  - efficiently learn relations of tokens
  - scale well on larger datasets
  - → achieve new state-of-the-art performance

Attention in Transformers



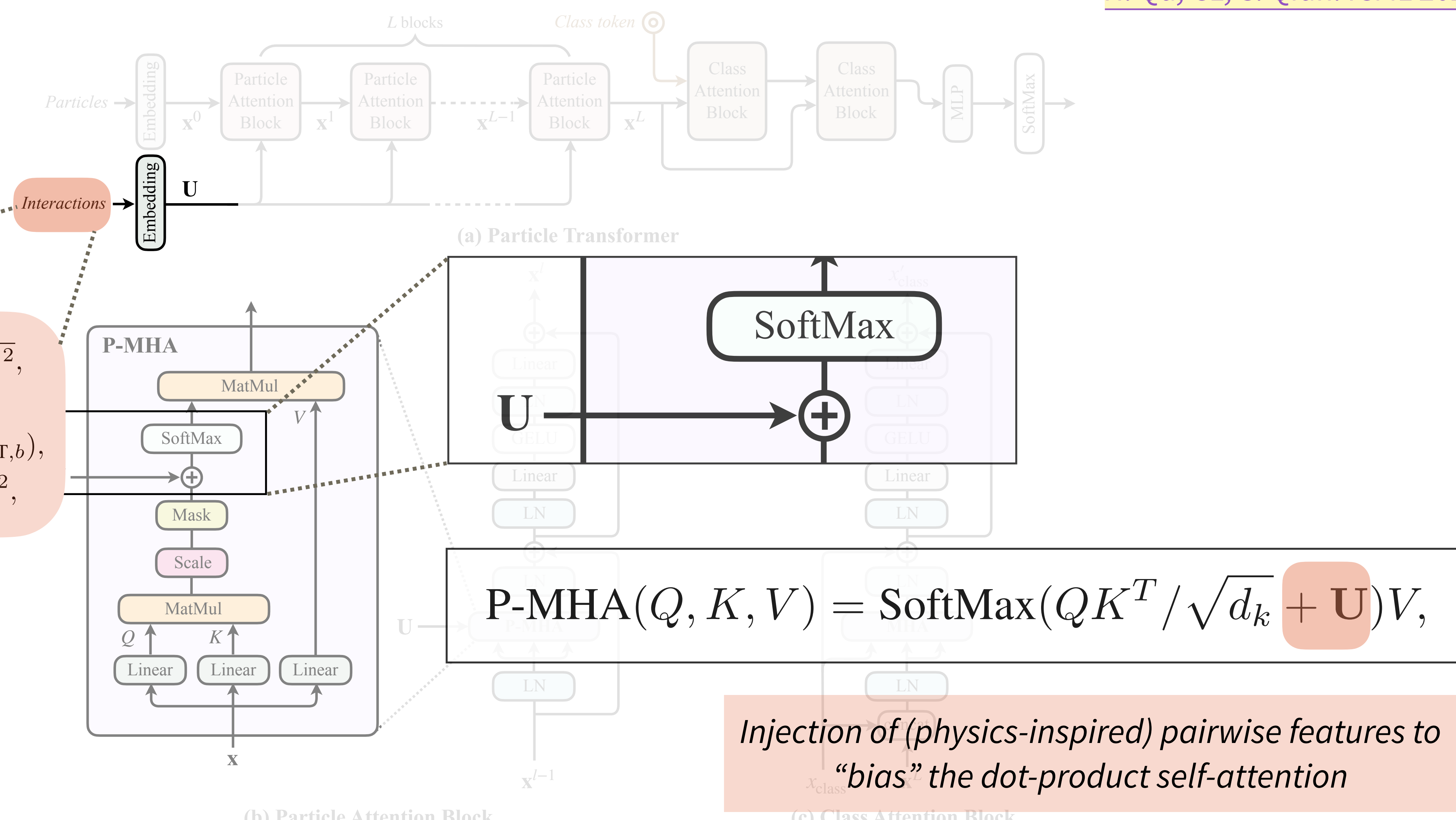
Each token (particle) talks to every other token

Same prototype across the fields

ParT: better adapt Transformers to jet data

- Transformer tailored for particle physics (e.g. jet tagging)
  - featuring its "attention bias" that **embed pairwise features respecting different levels of Lorentz symmetry**

H. Qu, CL, S. Qian. ICML 2022



$$\Delta = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2},$$
$$k_\mathrm{T} = \min(p_{\mathrm{T},a}, p_{\mathrm{T},b})\Delta,$$
$$z = \min(p_{\mathrm{T},a}, p_{\mathrm{T},b})/(p_{\mathrm{T},a} + p_{\mathrm{T},b}),$$
$$m^2 = (E_a + E_b)^2 - \|\mathbf{p}_a + \mathbf{p}_b\|^2,$$

and many other possible pairwise features…

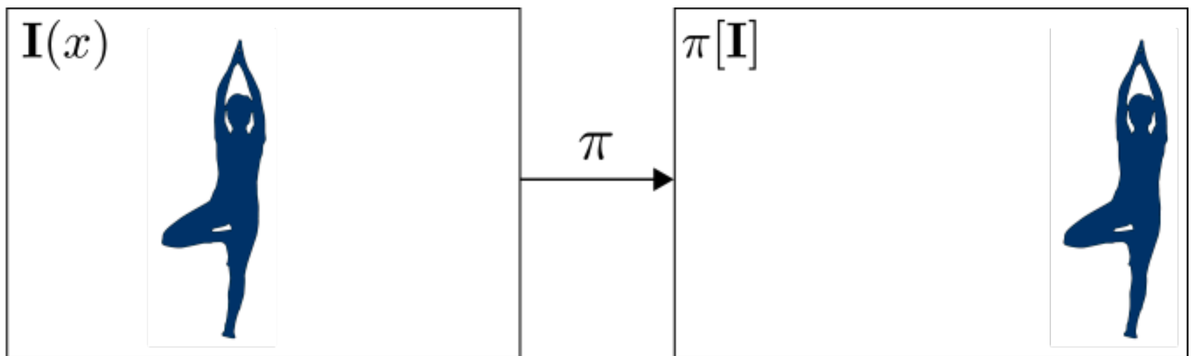
$$\text{P-MHA}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d_k} + \mathbf{U})V,$$

Injection of (physics-inspired) pairwise features to "bias" the dot-product self-attention

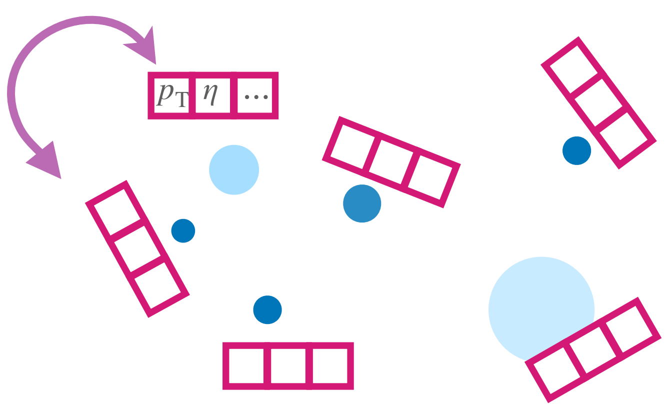
Backgrounds on symmetries and inductive biases

➔ Inherent symmetries of the dataset → inductive bias to improve NN performance

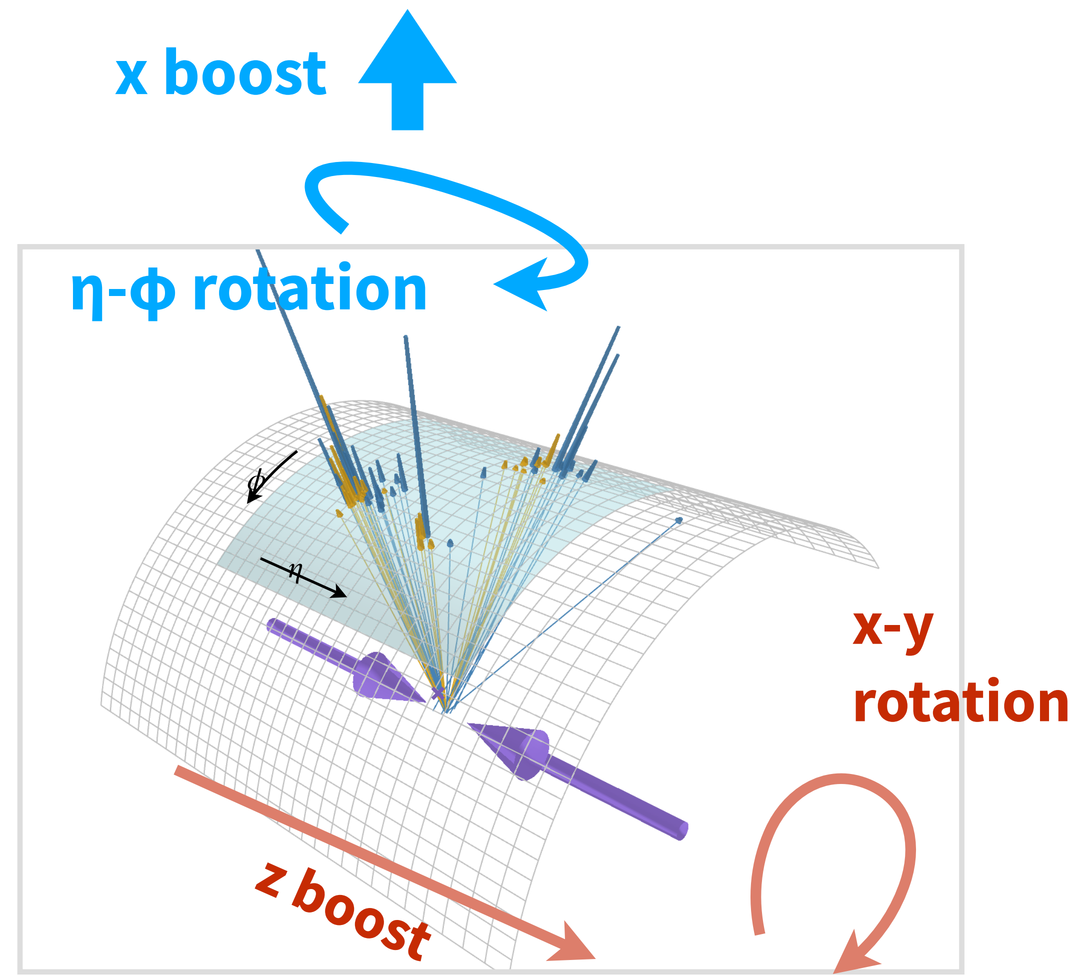
(CNN's advantage)



Translation (of image patches)



Permutation (of particle records)



Lorentz transformation

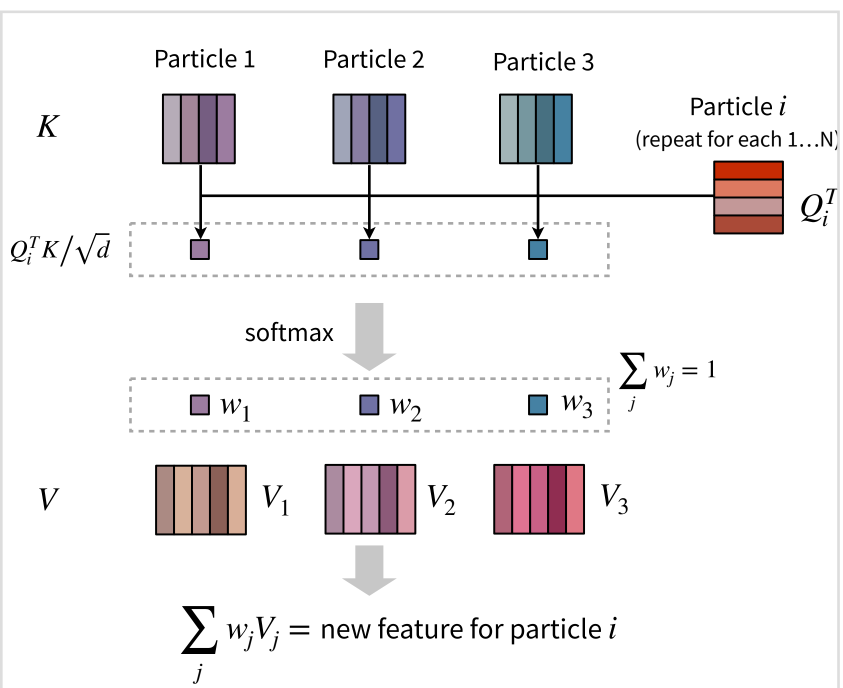
➔ Jets have symmetries under permutations & Lorentz transformations

Discussion in PRD 109, 056003 (2024)

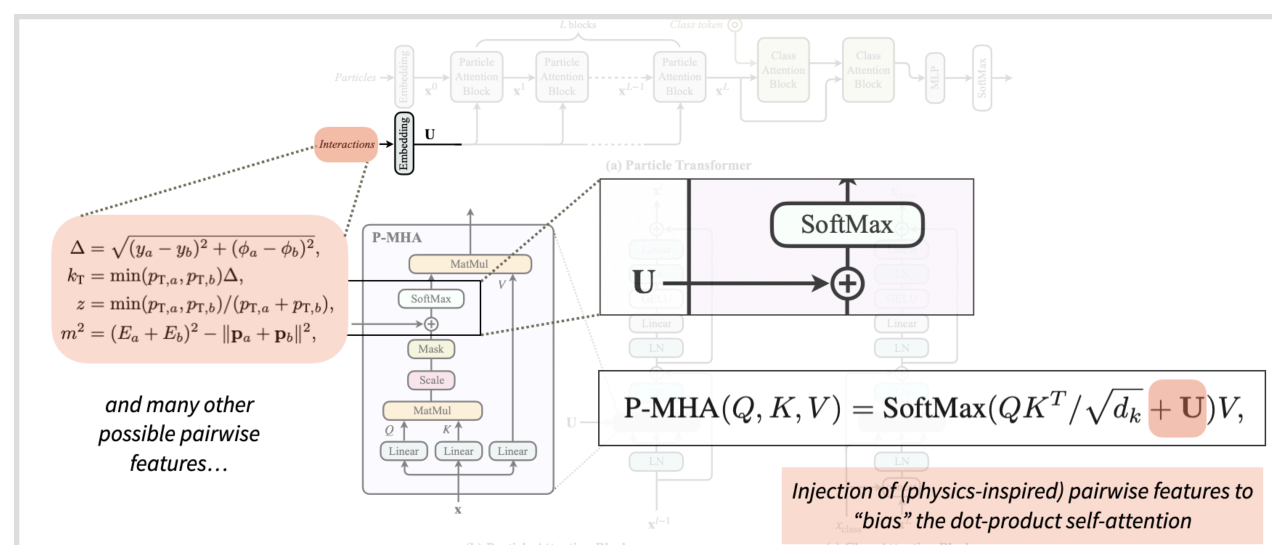
The ParT “engineering blueprint”

Plain Transformer + ***Inductive bias for particle-format data***

Particles as tokens



Permutation invariance: no particles’ positional embedding
Lorentz invariance: pairwise masses injected as attentive bias
(solution is close to AlphaFold)



Advances in Transformer models

1. Better scaling capability with model & dataset sizes

	All classes Accuracy	All classes AUC	$H \to b\bar{b}$ Rej$_{50\%}$	$H \to c\bar{c}$ Rej$_{50\%}$	$H \to gg$ Rej$_{50\%}$	$H \to 4q$ Rej$_{50\%}$	$H \to \ell\nu qq'$ Rej$_{99\%}$	$t \to bqq'$ Rej$_{50\%}$	$t \to b\ell\nu$ Rej$_{99.5\%}$	$W \to qq'$ Rej$_{50\%}$	$Z \to q\bar{q}$ Rej$_{50\%}$
ParticleNet (2 M)	0.828	0.9820	5540	1681	90	662	1654	4049	4673	260	215
ParticleNet (10 M)	0.837	0.9837	5848	2070	96	770	2350	5495	6803	307	253
ParticleNet (100 M)	0.844	0.9849	7634	2475	104	954	3339	10526	11173	347	283
ParT (2 M)	0.836	0.9834	5587	1982	93	761	1609	6061	4474	307	236
ParT (10 M)	0.850	0.9860	8734	3040	110	1274	3257	12579	8969	431	324
ParT (100 M)	0.861	0.9877	10638	4149	123	1864	5479	32787	15873	543	402

H. Qu, CL, S. Qian. ICML 2022

Dataset size scaled up

JetClass: dataset reaching **100 M** entries
- close to real experimental situations

performance improvements: ParT > ParticleNet

Model size scaled up

Larger ParT model to build real jet taggers in CMS (Global Particle Transformer, GloParT)

	ParT-lite	ParT-B	ParT-L (GloParT)
Input embed. dim.	(64, 256, 64)	(128, 512, 128)	(256, 1024, 256)
Pairwise feat. embed. dim.	(32, 32, 32, 8)	(64, 64, 64, 8)	(128, 128, 128, 16)
Transformer dim.	64	128	256
Number of heads	8	8	16
Fully-connected layer dim.	(512, 316)	(1024, 316)	(1024, 316)
Initial LR	6.75×10^{-3}	4×10^{-3}	2×10^{-3}
Batch size	768	512	256
Epochs	30	50	50

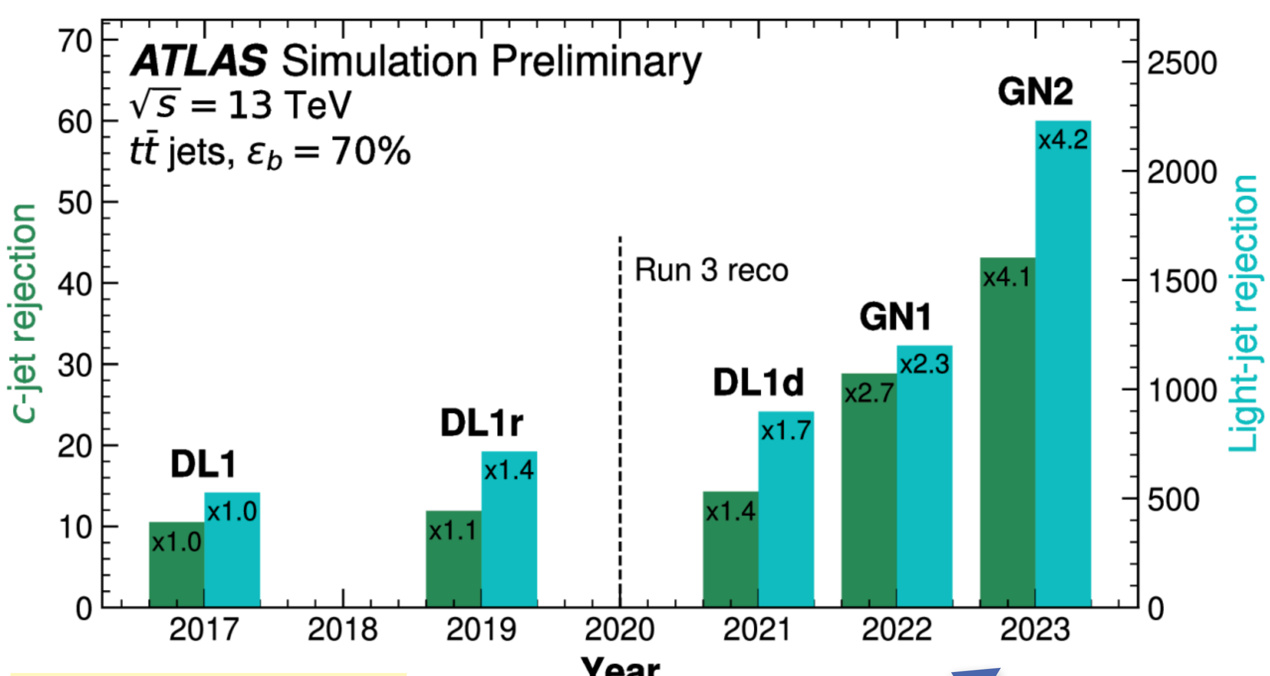
Advances in Transformer models

1. Better scaling capability with model & dataset sizes

➔ ATLAS/CMS "flagship" small-R jet taggers have all switched to the **Transformer architectures** (with training dataset size reaches o(100M) level)

- ❖ huge progress has been made from **2016 (early Run-2)** to **2024 (mid-Run3)** ! (rejection rate of c-jet & light-jet, for b-tagging)

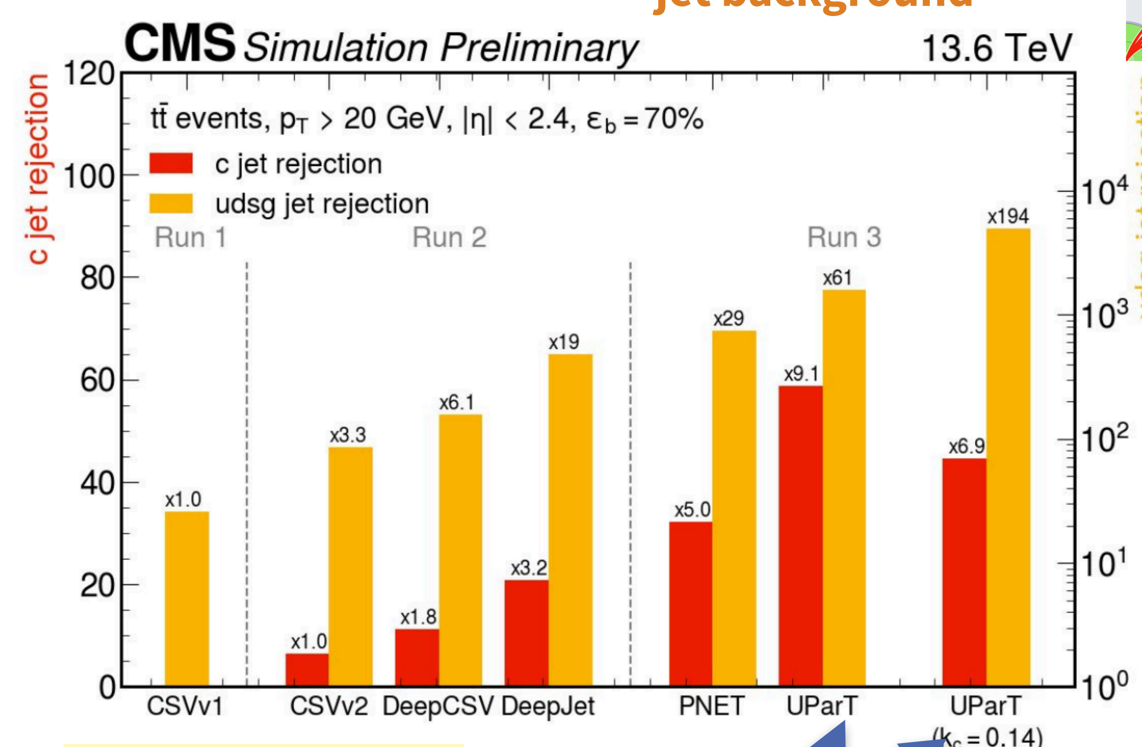
ATLAS EXPERIMENT



ATL-FTAG-2023-01

Latest ATLAS tagger for small-R jets: Transformer-based GN2

x50 rejection on light-jet background



CMS-DP-2024-066

Latest CMS tagger for small-R jets: Unified Particle Transformer (UParT)

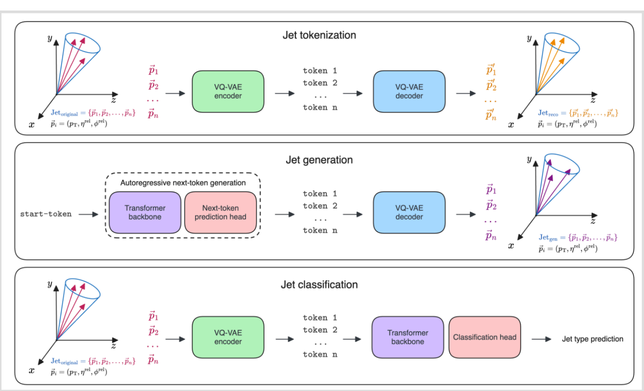
Advances in Transformer models

2. Building comprehensive / base / foundation HEP models

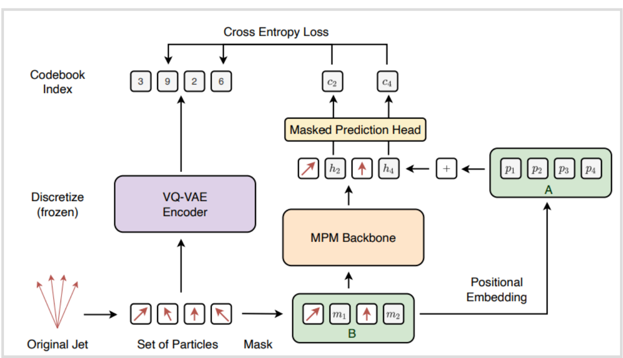
➔ **The ultimate goal: design a unified HEP model to analyze jets/events:**
- ❖ comprehensive phase space coverage
- ❖ one model handling all tasks - multimodality

➔ Engineering solutions:
- ❖ self-supervised learning to learn jet representations
- ❖ hybrid (multimodal) training across tasks: jet tagging, property regression, reconstruction/ generation…
- ❖ Model pre-training followed by “fine-tuning” to downstream tasks

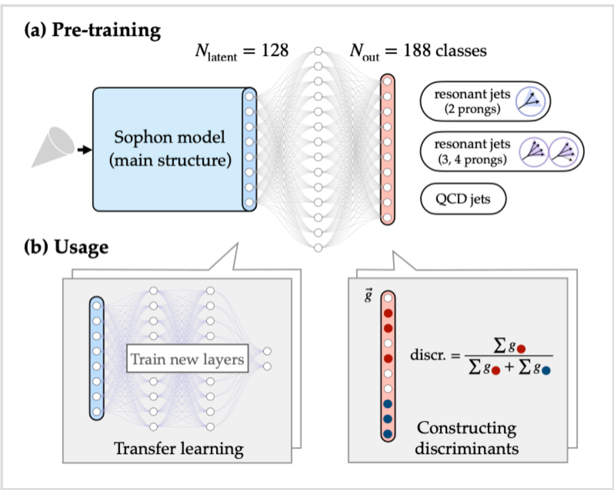
Recent work examples:



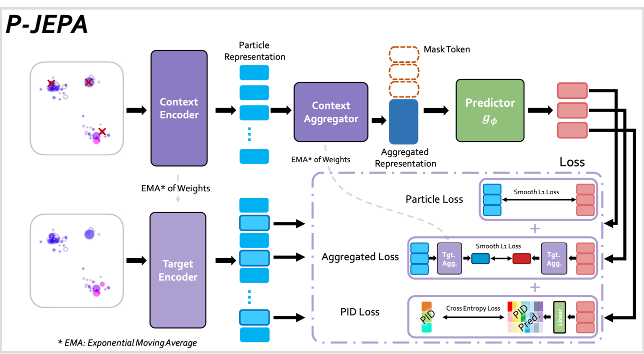
OmniJet-α
(GPT-like, next-token prediction to learn jet properties)
MLST, 5 035031 (2024)



Masked Particle Modelling
(SSL with Masked autoencoder (MAE) style)
2401.13537



Sophon model
(giant classifier for full jet phase space coverage)
2405.12972

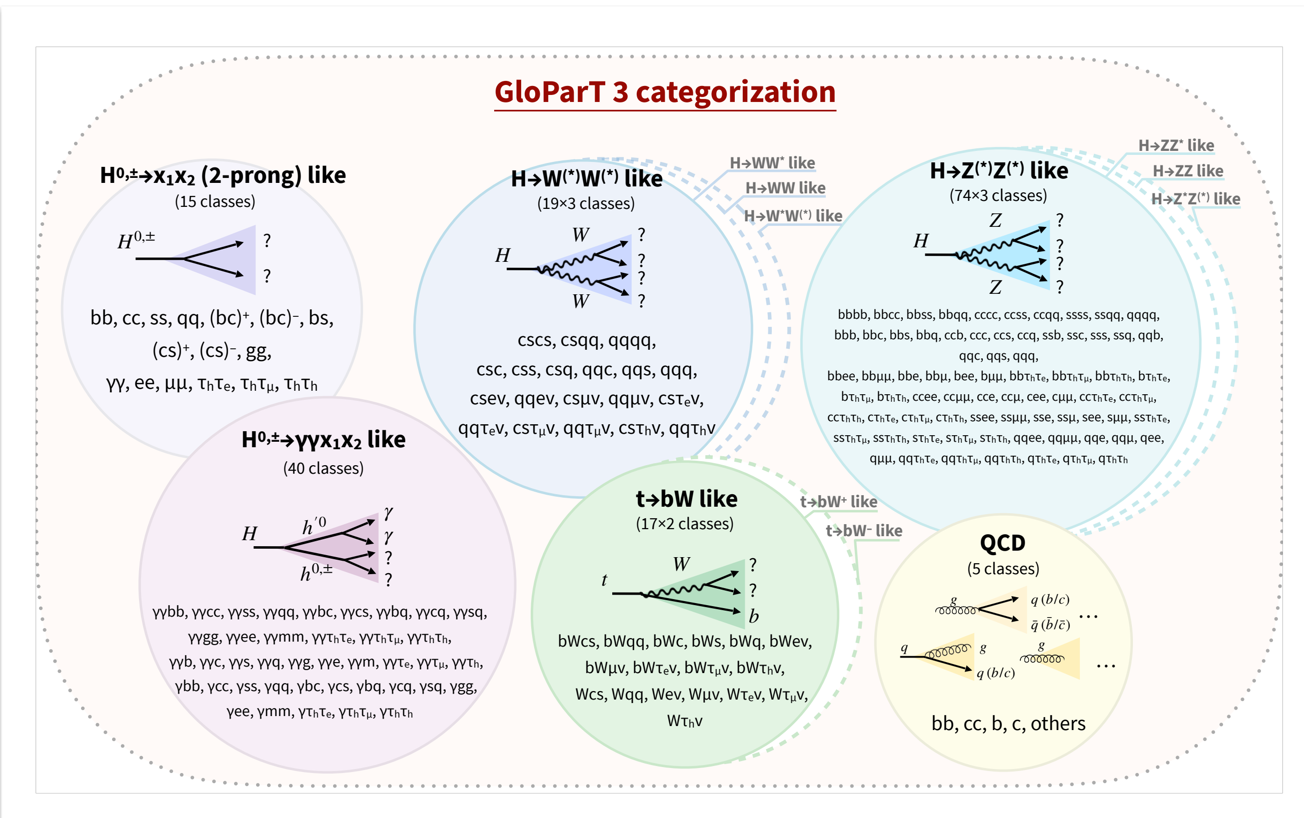


p-jepa
(jet embedding prediction)
see e.g. H. Qu’s talk

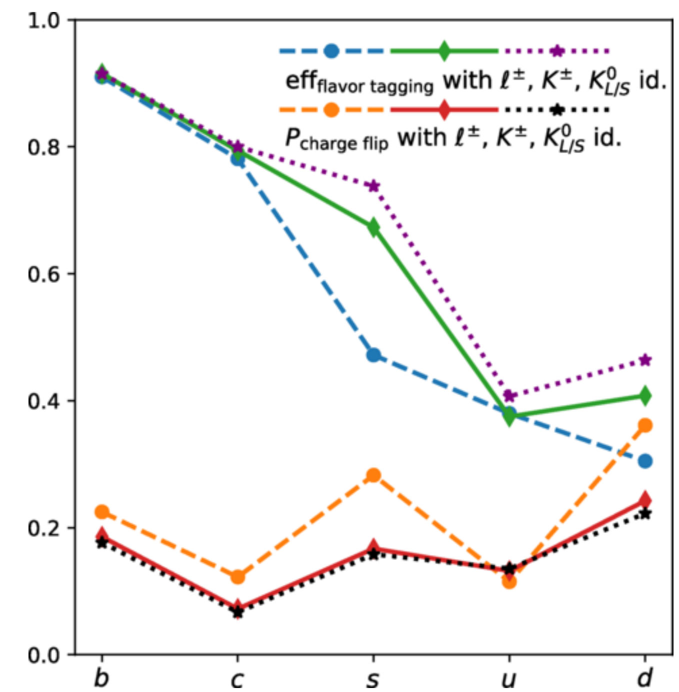
Advances in Transformer models

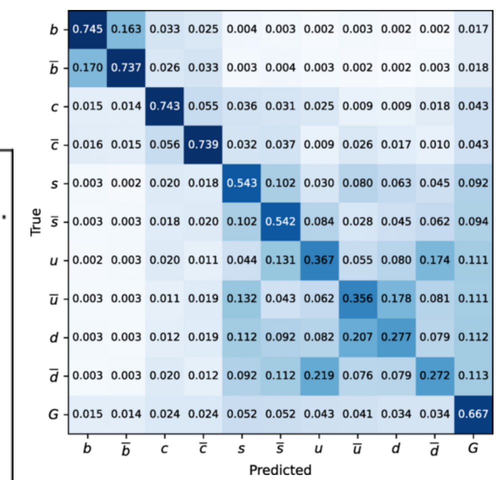
2. Building comprehensive / base / foundation HEP models

Mature experimental solutions



Global Particle Transformer (GloParT) in the CMS experiment (the giant jet model for tagging + mass regression)
- Sophon's CMS realization





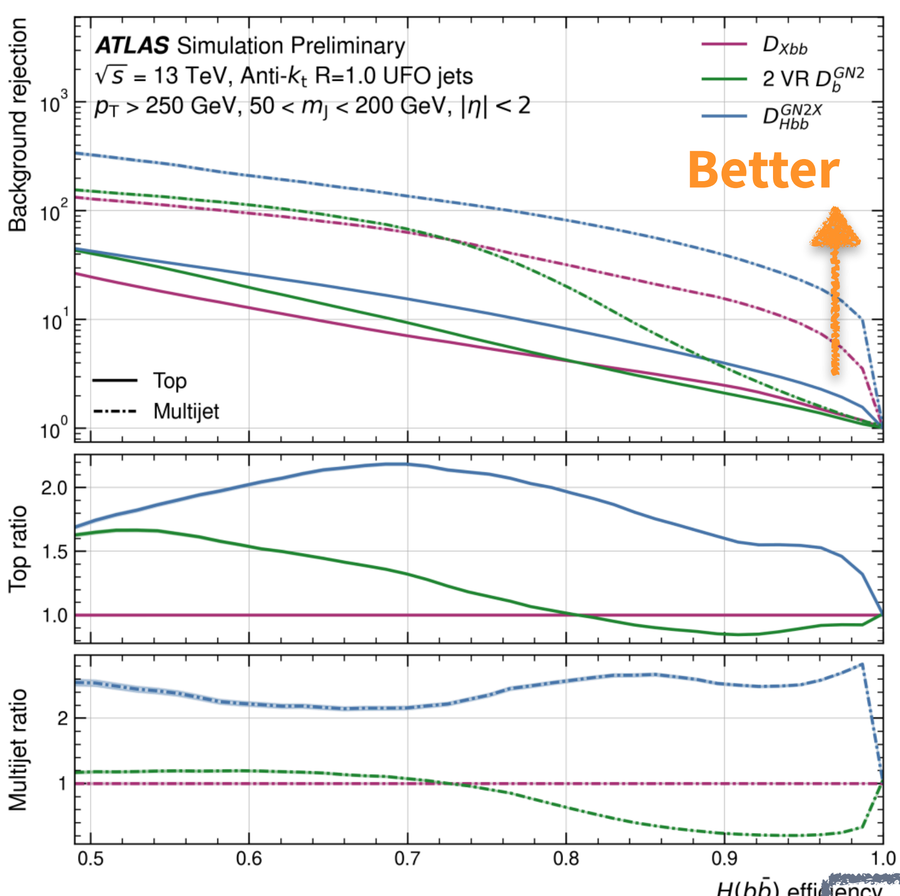
universal jet-origin identification solution (for all quark flavours and charges) for CEPC

H Liang, Y Zhu et al. PRL. 132, 221802

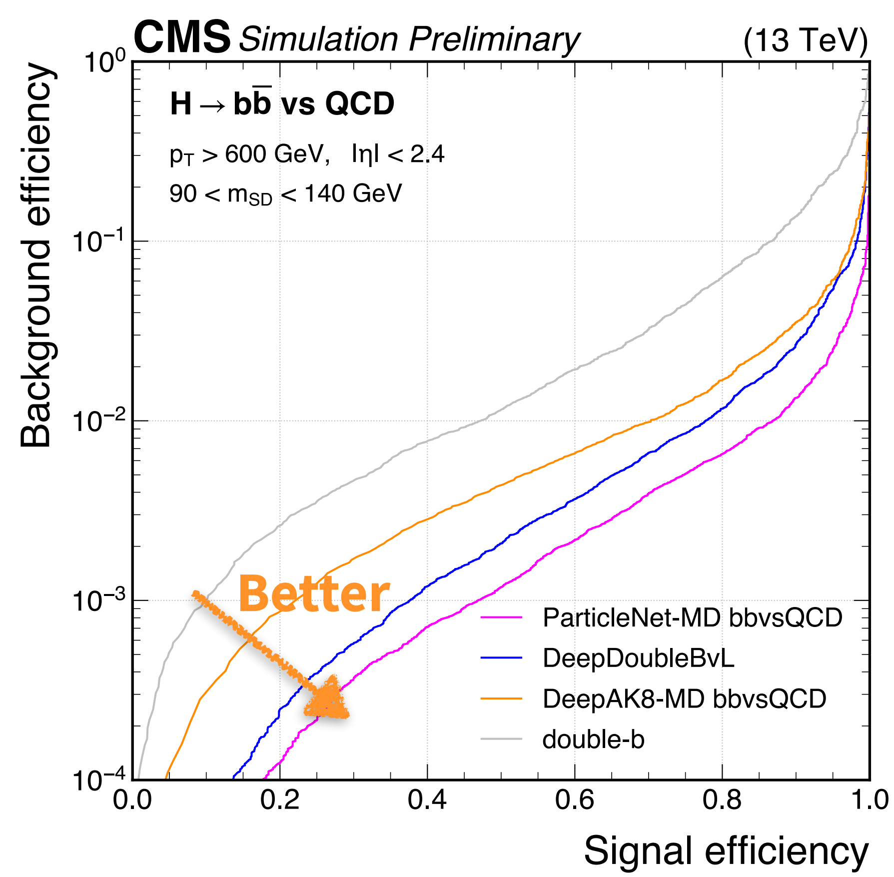
Future insights: boundary of jet identification?

- ➔ The statistical essence of classification via DNN is to let the network to fit the underlying pdf ratios: $\rho_A(x)/p_B(x)$
  - ❖ better DNN architectural design + training strategy → better estimation of pdfs
- ➔ We have seen consistent improvements over the past 5 years, but there is no sign that boundaries are reached
- ➔ Understanding the boundary is crucial! (e.g. 2411.02628)

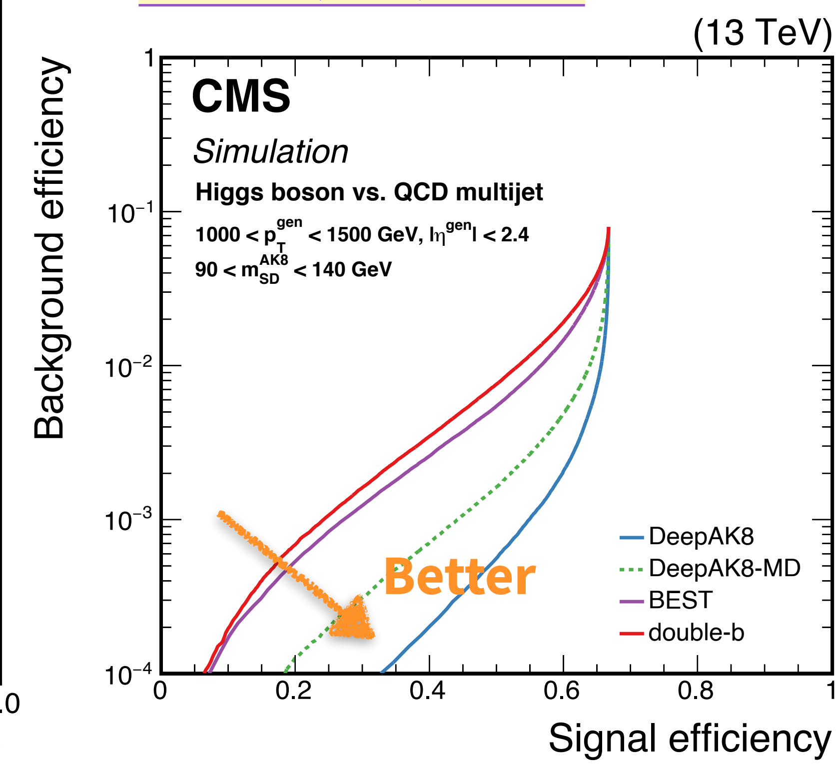
ATL-PHYS-PUB-2023-021



CMS-PAS-BTV-22-001



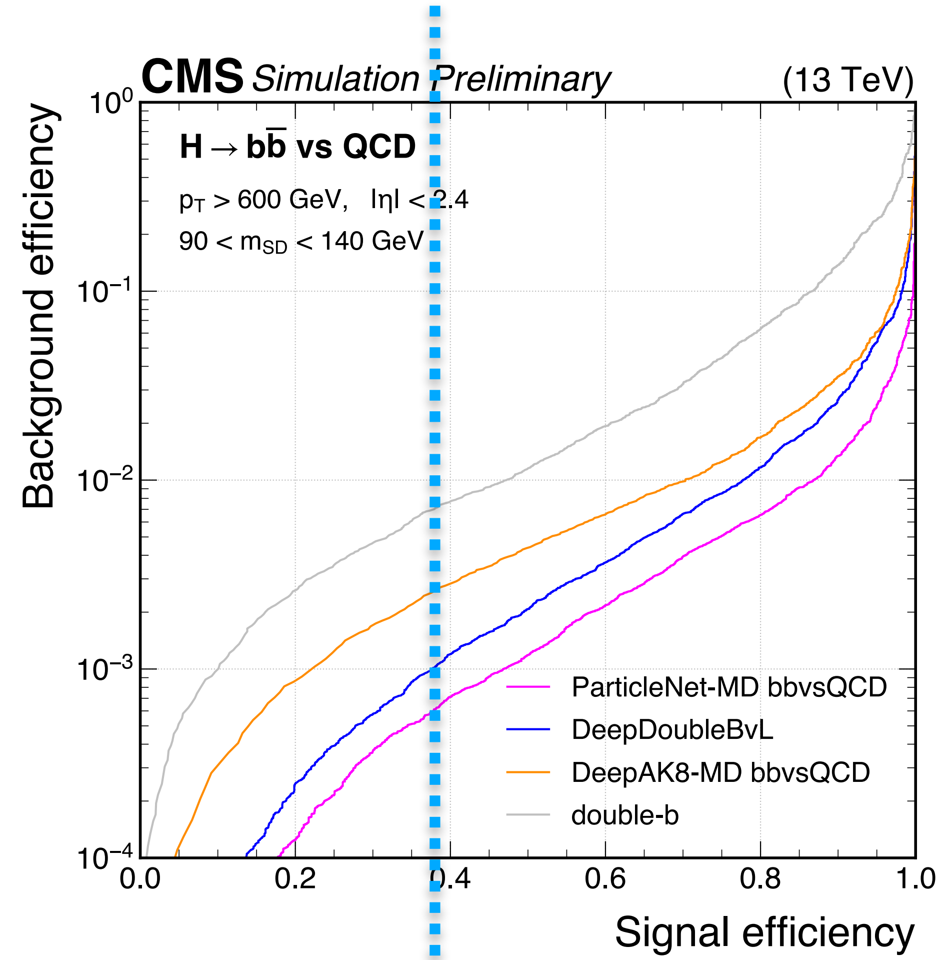
JINST 15 (2020) P06005



Consistent improvements seen; no boundaries reached

Future of analyzing hadronic events?

- ➔ Jet data / hadronic events are more complex objects to analyze than thought
  - ❖ not easy to touch the boundaries
- ➔ Small improvements have a large impact in the scientific result
  - ❖ popular metrics are classification accuracy/AUC, where usually small improvement is seen, but what is crucial is the "background rejection rate" $(1/\epsilon_B)$
  - ❖ i.e. at the working point of TPR (ϵ_S) ~ 0.5, but FPR (ϵ_B) ~ 1e-3
  - ❖ **FPR suppressed by ×2**
    → discovery sensitivity ×√2
- ➔ Capabilities to analyze hadronic-final-state processes (at the LHC) have been underestimated



Here is the working point of our concerns

Conclusion and outlook

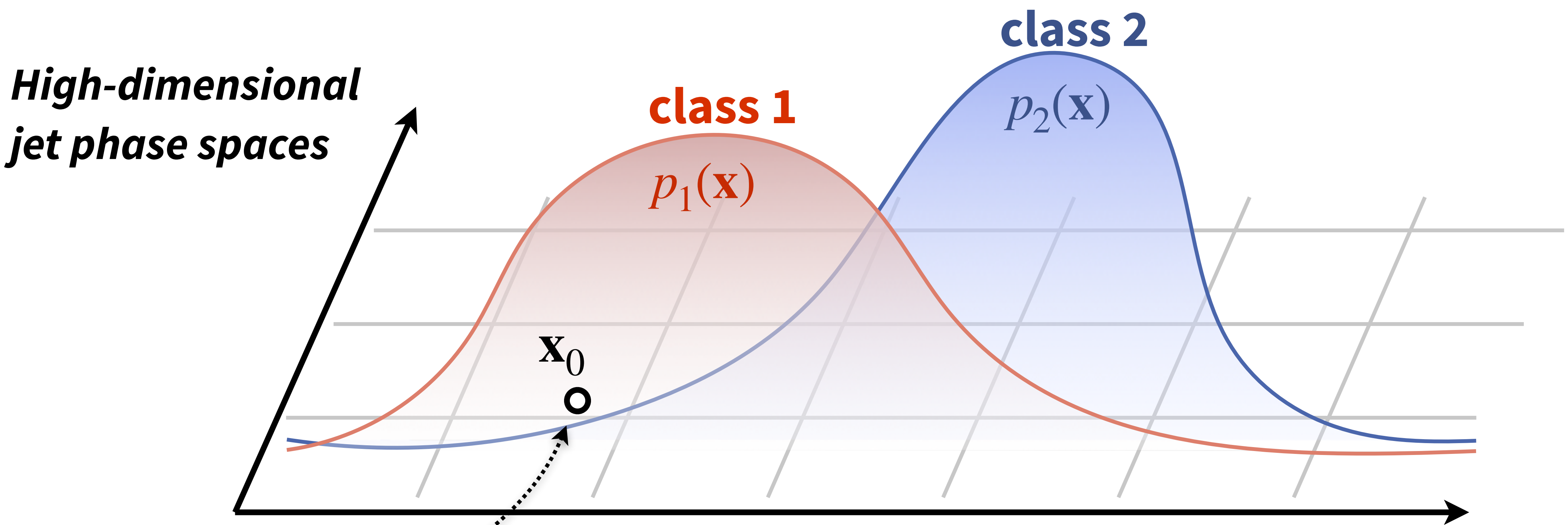
- ➔ Transformers have revolutionized the entire AI field, including their applications in HEP-ex and jet physics
  - ❖ jet tagging performance is brought to a new level
    - ParT is a baseline model (Transformer arch w/ proper inductive biasing)
    - engineering experiences are acquired and overviewed in this talk
- ➔ Next up?
  - ❖ Improving Transformers?
    - efficient Transformers (address the o(N^2) computation cost in self-attention)
    - better inductive biasing (e.g. relaxing pairwise embedding: L-GATr 2405.14806, 2411.00446; new embedding solution: MIParT CPC. 49 (2025) 1, 013110)
  - ❖ Better pre-training of jet Transformer models?
    - Current solutions are very open (self-/semi-/fully-supervised? variation of training targets)
      — always note that improving jet-analysis performance is the only criterion!
    - Need insights from the AI experts!

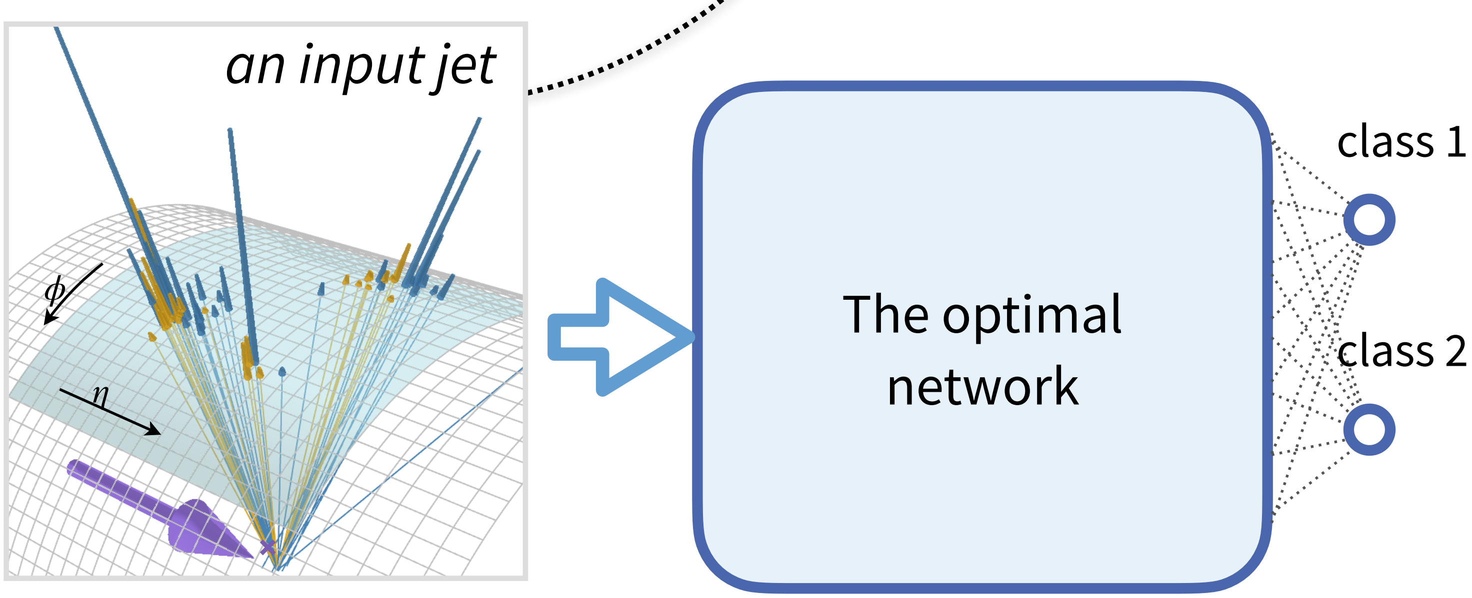
Backup

Statistical essence of jet tagging problem

➔ **Question: where is the limit of jet tagging?**

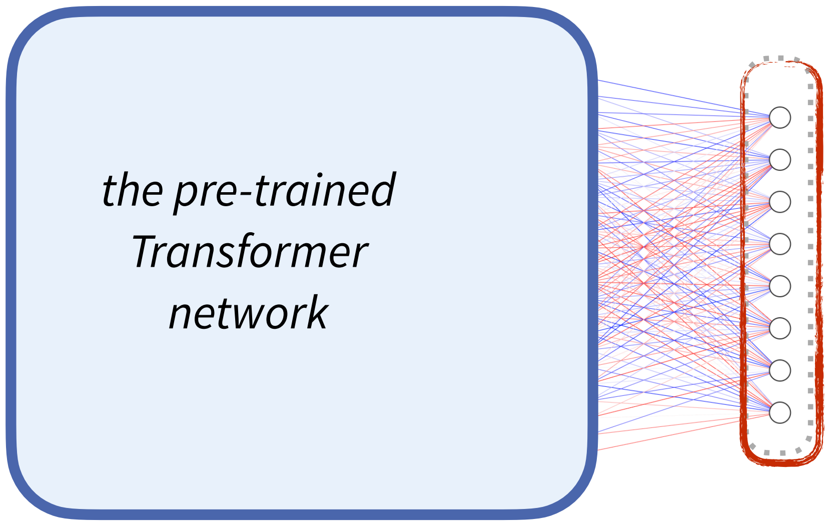
➔ Answer: the probability density ratio of two classes provides the optimal tagging

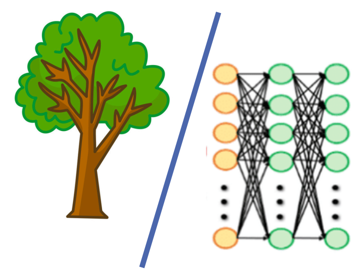




- Ideal classifier network results in $g_1 : g_2 : \ldots = p_1(\mathbf{x}_0) : p_2(\mathbf{x}_0) : \ldots$
- It is a direct estimation of p
- The **network capacity** decides how close the estimation is

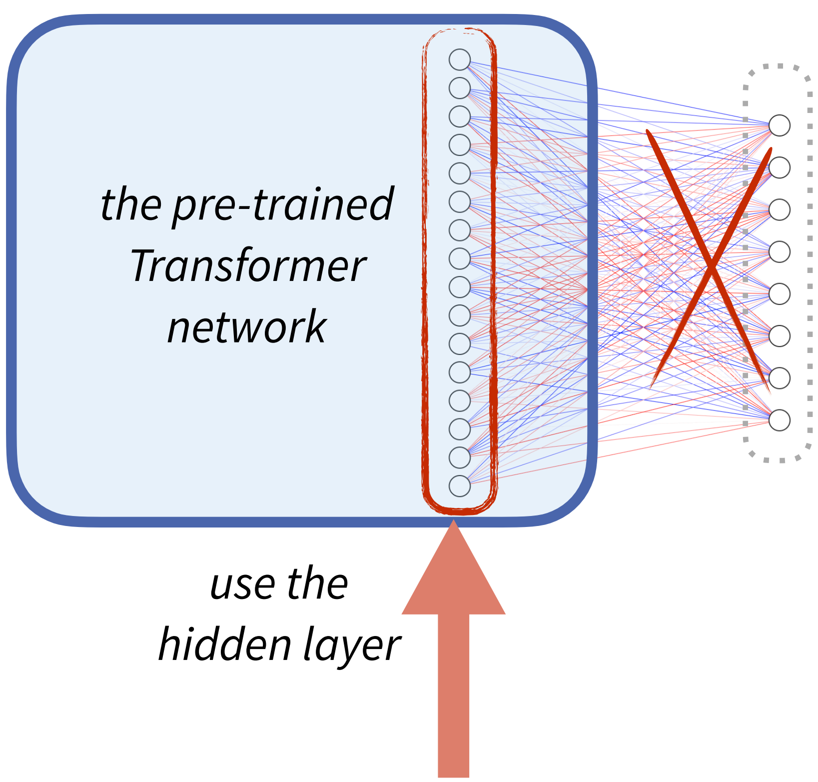
A glance into fine-tuning spirits

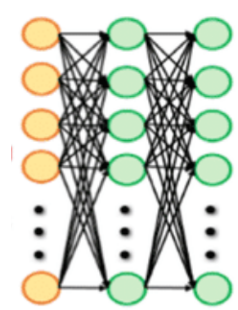




train a BDT or NN

customized scores!
(optimized for analysis)





train an NN

equivalently, this means to replace then retrain the last layer

customized scores!
(optimized for analysis)

This is a fine-tuning approach (specifically, transfer learning) in its equivalent form

CMS's path to develop Global Particle Transformer

Philosophy to develop Global Particle Transformer (GloParT) in CMS

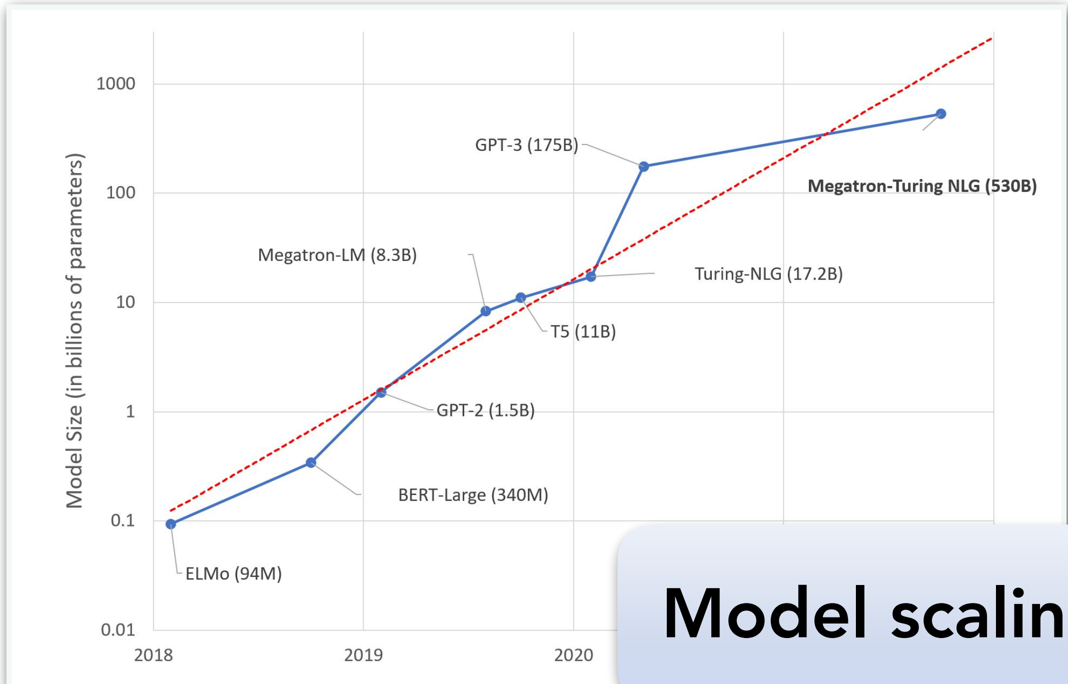
Good probability density estimators

- What is p? - the "differential cross section" of a process A on very high-dim space
- discriminating process A vs. B: estimate $p_A(\mathbf{x})/p_B(\mathbf{x})$ as best as we can
- need a model to **cover a variety of processes** $A, B, C, D, \ldots.$

$A \to BC$		e	μ	τ	q/g	b	t	γ	Z/W	H	$B =$ BSM
		$B =$ SM									
$C =$ SM	e	Z'	$\not R$	$\not R$	LQ	LQ	LQ	L^*	L^*	L^*	Many
	μ		Z'	$\not R$	LQ	LQ	LQ	L^*	L^*	L^*	
	τ			Z'	LQ	LQ	LQ	L^*	L^*	L^*	
	q/g				Z'	W'	T'	Q^*	Q^*	Q'	
	b					Z'	W'	Q^*	Q^*	B'	
	t						Z'	Q^*	T'	T'	
	γ							H	H	Z_{KK}	
	Z/W								H	$H^\pm/A$	
	H									H	
$C =$ BSM		Consider just the di-object search for resonant A → B C									Many

J.Kim *et al.* JHEP 04 (2020) 30 1907.06659

Generalization ability



Model scaling up

- one upstream pre-training, broad downstream applicability